Related Entity Finding Based on Co-Occurrence

Marc Bron Krisztian Balog Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl/

Abstract: We report on experiments for the Related Entity Finding task in which we focus on only using Wikipedia as a target corpus in which to identify (related) entitities. Our approach is based on co-occurrences between the source entity and potential target entities. We observe improvements in performance when a context-independent co-occurrence model is combined with context-dependent co-occurrence models in which we stress the importance of the expected relation between source and target entity. Applying type filtering yields further improvements results.

1 Introduction

The start of a new track usually means the introduction of a new task—in this case, related entity finding (REF)—to be solved in the absence of training data and a standard system design. In approaching such a task, a sensible strategy is to start with a general system design and subsequently extend and refine it. We investigate an approach based on co-occurrences of potential target entities with the source entity given in the topic statement. We consider two variants: a purely co-occurrence based model and a combination of this with a context dependent model that takes documents (in which both entities co-occur) in consideration as context. On top of this we experiment with applying a type filtering component. Our overal system design has the following components:

- Named entity recognition
- Named entity normalization
- (Context-independent) co-occurrence modeling
- Context-dependent co-occurrence modeling
- Type filtering
- Home page finding.

For the homepage finding part of the task we focus on the pipeline design; we decide on methods to use for named entity recognition (NER), named entity normalization (NEN), and homepage finding as well as how to combine these with a co-occurrence and type filtering component. As the components are mutually dependent and the evaluation is end to end, there is a risk of noise accumulating throughout the system, resulting in poor performance. So for the optional Wikipedia field we employ a different strategy and focus on the co-occurrence component, while minimizing the influence of other components in two ways: (i) NER and NEN are handled by considering Wikipedia as a repository of (normalized) known entities and (ii) homepage finding is handled by mapping entities to Wikipedia pages.

Our TREC 2009 submissions were plagued by a number of bugs. The homepage part of our runs achieves disappointing results. An analysis reveals two causes. First, a standard tagger is unsuitable for NER as it is too liberal in accepting strings as entities, thus polluting the set of candidate entities. Second, the homepage finding task is a difficult problem and our ad hoc solution (cf. Section 2.4.2) turns out to be unsuitable. As there is no value in analyzing these results any further, we leave this part as is and instead discuss our runs only considering the Wikipedia field, i.e., only using Wikipedia as the target corpus in which to identify relavant entities.

We find that considering only Wikipedia pages as entities overcomes the NER and homepage finding weaknesses in the REF pipeline. Through analysis of the co-occurrence component we find that combining the pure co-occurrence and the context dependent model improves over a pure co-occurrence model alone, and that type filtering further improves these results.

In this paper we report on the repaired runs, only using Wikipedia as the target corpus. We describe our approach in Section 2, our results in Section 3, and conclude in Section 4.

2 Approach

We formulate the entity ranking problem as follows. The goal is to rank candidate entities (e) according to $P(e|E,T,R)$, where E is the source entity, T is the target type, and R is the relation described in the narrative.

Instead of estimating this probability directly, we use Bayes' rule and reformulate it into:

$$P(e|E,T,R) \quad = \quad \frac{P(E,T,R|e) \cdot P(e)}{P(E,T,R)}. \qquad (1)$$

Next, we drop the denominator as it does not influence the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Related Entity Finding Based on Co-Occurrence	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Amsterdam,Intelligent Systems Lab Amsterdam (ISLA),Science Park 107,1098 XG Amsterdam, The Netherlands,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

ranking of entities, and derive our final ranking formula as follows:

$$\begin{aligned} & P(E,T,R|e)\cdot P(e) \\ & \quad = P(E,R|e)\cdot P(T|e)\cdot P(e) && (2) \\ & \quad = P(E,R,e)\cdot P(T|e) \\ & \quad = P(R|E,e)\cdot P(E,e)\cdot P(T|e) \\ & \quad = P(R|E,e)\cdot P(e|E)\cdot P(E)\cdot P(T|e) && (3) \\ & \quad \stackrel{\text{rank}}{=} P(R|E,e)\cdot P(e|E)\cdot P(T|e) && (4) \end{aligned}$$

In (2) we assume that the type is independent of the source entity E and the relation R. Next, we rewrite $P(E,R|e)$ to $P(R|E,e)$ so that it expresses the probability that relation R is generated by the two (co-occurring) entities (e and E). Finally, we rewrite $P(E,e)$ to $P(e|E)\cdot P(E)$ in (3) as the latter is a more convenient form for estimation, and we drop $P(E)$ in (4) as it does not influence the ranking (for a fixed source entity E). Given equation (4) we are left with the following components:

- $P(e|E)$: pure co-occurrence model,
- $P(R|E,e)$: context dependent model, and
- $P(T|e)$: type filtering.

In the following sections we describe our estimation methods for these components. In Section 2.4 we give a short overview of the other components of the pipeline.

2.1 Pure co-occurrence model

We use this component to express the strength of associations between the source entity and candidates, without considering the nature of their relation. We use pointwise mutual information as an estimate for $P(e|E)$:

$$P(e|E) = \frac{PMI(e,E)}{\sum_{e'} PMI(e',E)}$$

and $PMI(e,E)$ is defined as follows:

$$PMI(e,E) = \log \frac{c(e,E)}{c(e)\cdot c(E)},$$

where $c(e,E)$ is the number of documents in which e and E co-occur and $c(e)$ is the number of documents in which e occurs.

2.2 Context-dependent model

In this component we model the relations between the source entity and candidate target entities. We represent the relation between a pair of entities by a co-occurrence language model (θ_{Ee}), a distribution over terms taken from documents in which the source and candidate entity co-occur. By assuming independence between the terms in the relation R we arrive at the following estimate for this component:

$$P(R|E,e) = P(R|\theta_{Ee}) = \prod_{t\in R} P(t|\theta_{Ee})^{n(t,R)}, \qquad (5)$$

where $n(t,R)$ is the number of times t occurs in R. To estimate the co-occurrence language model θ_{Ee} we aggregate term probabilities from documents in which the two entities co-occur:

$$P(t|\theta_{Ee}) = \frac{1}{|D_{Ee}|} \sum_{d\in D_{Ee}} P(t|\theta_d), \qquad (6)$$

where D_{Ee} denotes the set of documents in which E and e co-occur and $|D_{Ee}|$ the number of these documents. $P(t|\theta_d)$ is the probability of term t within the language model of document d:

$$P(t|\theta_d) = \frac{n(t,d)+\mu\cdot P(t)}{\sum_{t'} n(t',d)+\mu}, \qquad (7)$$

where $n(t,d)$ is the number of times t appears in document d, $P(t)$ is the collection language model, and μ is the Dirichlet smoothing parameter, set to the average document length in the collection.

2.3 Type detection

The final component is used to filter entities by type. In order to perform type filtering we exploit the Wikipedia category structure; we map each of the (source) entity types ($T \in \{PER, ORG, PROD\}$) to a top category ($\text{cat}(T)$), e.g., "living people" and we create a similar mapping for entities to categories ($\text{cat}(e)$). With these two mappings we estimate $P(T|e)$ as follows:

$$P(T|e) = \begin{cases} 1 & \text{if } \text{cat}(e) \cap \text{cat}^{L_n}(T) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

We also perform category expansion for entity types by adding direct child categories to each level and write $\text{cat}^{L_n}(T)$, where L_n is the chosen level of expansion. For example the second level L_2 contains the top categories (of level L_1) and all direct child categories.

2.4 The rest of the pipeline

The remaining components of the REF pipeline, i.e., named entity recognition and normalization as well as homepage and Wikipedia page finding, are described below.

2.4.1 Entity Recognition and Normalization

On Clueweb Category B we use the Stanford named entity tagger to recognize entities (Finkel et al., 2005). The tagger recognizes 4 entity types: person, organization, location, and miscellaneous.

On Wikipedia we handle named entity recognition by only considering anchor texts from links within Wikipedia as entity occurrences. We obtain an entity's name by removing the Wikipedia prefix from the anchor URL.

For NEN we map URLs to a single entity variant. Here we make use of Wikipedia redirects that map common alternative spellings or references (e.g., "Schumacher," "Schumi" and "M. Schumacher") to the "canonical form" of an entity ("Michael Schumacher").

2.4.2 Homepage and Wikipage finding

Once we have obtained a ranked list of entity names, we submit a query "official homepage of ⟨ENTITY⟩" for each to obtain a list of documents. To determine if a document is a homepage we use edit distance between a documents URL and the entity name and use the highest scoring documents as homepages.

For matching entities to Wikipedia pages we use the anchor URL and return the corresponding target destination; the entity's Wikipedia page.

3 Results

The runs we focus on are centered around the co-occurrence component; ilpsEntBL and ilpsEntem. In our original runs the Wikipedia fields were not included, due to a bug in our code. As our focus is now solely on Wikipedia, we have generated new runs and replaced all homepage (HP) fields by a dummy document ID. We also continue experiments with the level of category expansion for our type filtering component and vary the levels from no filtering (L_0) to L_2, L_4 and L_6.

Table 1: Total score for each of our Wikipedia based runs.

runID	nDCG_R	P10	pri_ret	rel_ret
ilpsEntBL_L0	0.0204	0.0100	11	**23**
ilpsEntBL_L2	0.0325	0.0350	44	2
ilpsEntBL_L4	0.0266	0.0300	35	3
ilpsEntBL_L6	0.0227	0.0100	29	6
ilpsEntem_L0	**0.0657**	**0.0650**	58	1
ilpsEntem_L2	0.0616	0.0650	**69**	14
ilpsEntem_L4	0.0540	0.0550	64	6
ilpsEntem_L6	0.0575	0.0600	68	10

In order to compare our runs we use the number of primary Wikipedia pages (pri_ret), where primary means the encyclopedic entry of an entity, normalized discounted cumulative gain (nDCG), precision at 10 (P@10) and the number of relevant Wikipedia pages (rel_ret).

ilpsEntBL combines the pure co-occurrence model with type filtering:

$$\text{score}(e) = P(e|E) \cdot P(T|e)$$

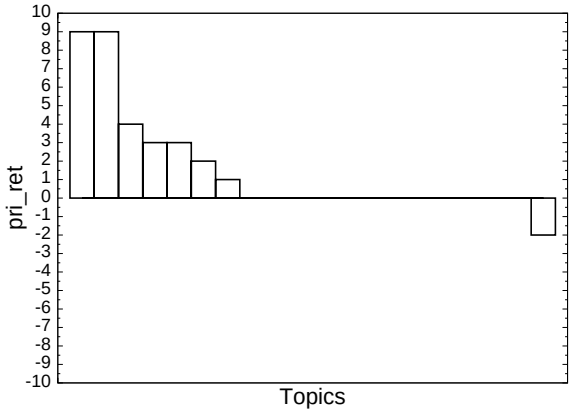

Figure 1: Difference in the number of Wikipedia pages (pri_ret) found by the pure co-occurrence model and the combination with the context dependent model. A positive value indicates that more Wikipedia pages are found when the models are combined.

ilpsEntem combines the pure co-occurrence model with the context dependent model and type filtering.

$$\text{score}(e) = P(R|E,e) \cdot P(e|E) \cdot P(T|e)$$

Table 1 shows the results for the Wikipedia only runs. We observe that the model that combines context and pure co-occurrence outperforms the pure co-occurrence model in all runs. The influence of different levels of type filtering on the pure co-occurrence model shows a clear trend; less expansion improves results. In the combined model the differences are smaller, suggesting that context reduces the number of non relevant entities of the wrong type in the top of the ranking. Figure 1 shows the difference between the number of primary pages found by each of the models per topic (filtering level 4). A positive value indicates that more Wikipedia pages are found when the models are combined. We observe that only on topic 10 less primary pages are found, on 7 topics using context increases that number and on 13 topics context does not influence the number of primary Wikipedia pages found.

Our context dependent model finds reasonable numbers of primary pages. The P@R and nDCG_R scores, however, are low. Topic 17 (i.e., E: "The food network", R: "Chefs with a show on the food network" and T: "person") is a good example of a topic that achieves good recall and poor P@10 and nDCG scores. Table 2 shows the top 10 entities returned for topic 17 and their frequencies. We observe that the frequencies of the top 5 entities returned by both models are very

Pure co-occurrence		
Rank	Entity name	Frequency
1	Wayne Harley Brachman	5
2	Kerry Vincent	1
3	Jacqui Maloufa	5
4	Glenn Lindgren	3
5	Geof Manthorne	2
Context dependent		
Rank	Entity name	Frequency
1	Gennaro Contaldo	10
2	Asako Kishi	18
3	Yutaka Ishinabe	13
4	Alpana Singh	15
5	Masahiko Kobe	16
34	**Anne Burrell**	16
53	**Robert Irvine**	63
75	**Tyler Florence**	83
82	**Cat Cora**	99
87	**Michael Symon**	80

Table 2: Entities returned for topic 17 by the pure co-occurrence model (top) and the context dependent model (bottom). Relevant entities are indicated in bold.

low. On the other hand, the relevant entities (indicated in bold face) are more frequent and also ranked lower. It turns out that the use of PMI in our pure co-occurrence model creates a bias towards entities that occur less frequent. This is an inherent property of PMI as is noted in Manning and Schuetze (1999) and indicates that we need to consider alternative co-occurence statistics to obtain high precision on the REF task.

4 Conclusion

In our participation this year we set out to design a related entity finding system and to investigate the applicability of co-occurrence based models to the REF task. For our main homepage finding run we focused on identifying and assembling components into a REF system. The NER tool and homepage finding method, however, turned out to be unsuitable and resulted in disappointing results. The availability of this years topics as training set will facilitate developing a more robust REF system and should help eliminate issues of this kind in the future.

For our Wikipedia runs we eliminated interfering components as much as possible. We removed noise introduced by NER by only considering anchor URLs as entities and homepage finding by mapping entities to Wikipedia pages. This allowed us to focus on the co-occurrence and type filtering components of our system. We found that using PMI for the pure co-occurrence model produces a bias towards infrequent entities, suggesting the need for other estimation methods. When the pure co-occurrence model is combined with contextual information results improve on all runs and on all but one topic. This suggests that context is either of use for REF or does not influence the result.

Our P@10 and nDCG_R scores are low, a fact caused by the use of PMI in our pure co-occurrence model. In future work we plan to investigate other estimation methods for this model and to construct a more effective REF pipeline by evaluating various methods and tools for the NER and homepage finding components.

5 Acknowledgments

This research was supported by the DAESO and DuOMAn projects carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (`http://www.stevin-tst.org`) under project number STE-05-24 and STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 640.001.-501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

6 References

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Manning, C. D. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.